Recursos para el Procesamiento de Lenguaje Natural

Ciencia de Datos y Lenguaje Natural

Grupo PLN - INCO

Universidad de la República

Tipos de recursos

- El dominio de una lengua implica importantes conocimientos acumulados.
- Se conoce el repertorio de palabras y cómo se asocian a conceptos: léxico.
- Se conocen (consciente o inconscientemente) reglas de ordenamiento y modos de componer significados: sintaxis.
- La comunicación implica significados, significado "literal", significado asociado al uso social, relaciones entre significados.

Tipos de recursos

- Se requieren contrapartes informáticas para todos los tipos de recursos.
- Ya vimos ejemplos del software que maneja el repertorio léxico, tiene el *know how* necesario para realizar el análisis sintáctico, y sabe reconocer y clasificar entidades con nombre. Algunos ejemplos son Freeling, spaCy, stanza.

Recursos-semántica léxica

- Enfoque analítico. Conceptos y relaciones entre ellos. Un ejemplo es WordNet. Hecho a mano, conceptual. Registra distintas acepciones de las palabras. Registra relaciones entre las acepciones (más general/más específico, sinónimos, antónimos, parte/todo)
- Semántica distribuida. Refleja como se usan las palabras. Construido a partir de un corpus.

Recursos-conocimiento del mundo

- Existen además varias bases que refieren a la vez a hechos del mundo y a organización de los conceptos. Una mezcla entre esquema conceptual y bases de datos fácticas.
- Se las suele llamar Grafos de Conocimiento (Knowledge Graphs), término introducido por Google para un recurso privado generado inicialmente a partir de Freebase.
- Veremos algunos casos: CYC, YAGO y Wikidata

CYC - enCYClopedic knowledge

- Proyecto AI de larga data (desde el 1984)
- Contiene una ontología y una base de conocimiento de amplio espectro.
- Apunta a capturar ¸conocimiento de sentido común", presente en el contexto de la comunicación. Se usa permanentemente de modo implícito.
- ejemplos:
  - los padres quieren y protegen a sus hijos
  - el limón es ácido
  - de noche hay oscuridad

CYC - Ontología

- La ontología llegó a 100,000 términos en el 1994, y en el 2017 a 1,500,000 términos. Incluye:
- 416,000 colecciones (tipos, clases, clases naturales, con tipos de cosas tales como pez y tipos de acciones tales como pescar)
- 42,500 predicados (relaciones, atributos, propiedades, funciones),
- Cerca de un millón de entidades muy conocidas tales como TheUnitedStatesOfAmerica, BarackObama, TheSigningOfTheUSDeclarationOfIndependence, etc.

CYC - Base de Conocimiento

- La base de conocmiento se hizo a mano
- cerca de un millón de ítems en 1994, 24.5 millones en 2017
- Se estima que llevó más de 1000 años hombre construir lo que hay hasta ahora
- Se ha tratado de mantener reducida en tamaño la base de conocimiento: p.ej., no se agregan enunciados que se puedan deducir por la clausura transitiva de otros

CYC - ejemplo

(isa BurningOfPapalBull SocialGathering)
Because this is an instance of SocialGathering, it is an known to be an instance of Event and to have attendees.

(eventOccursAt BurningOfPapalBull
CityOfWittenburgGermany)

(dateOfEvent BurningOfPapalBull
(DayFn 10 (MonthFn December (YearFn 1520))))

(attendee BurningOfPapalBull
MartinLuther-ReligiousFigure)
Martin Luther is already represented in the KB, along with basic biographical information such as birth and death date, country of residence, and native language.

(relationInstanceExistsMin BurningOfPapalBull
attendees UniversityStudent 40)
At least forty university students attended the event. RelationInstanceExistsMin is a rule macro predicate.

(isa BurningOfPapalBull-Document CombustionProcess)

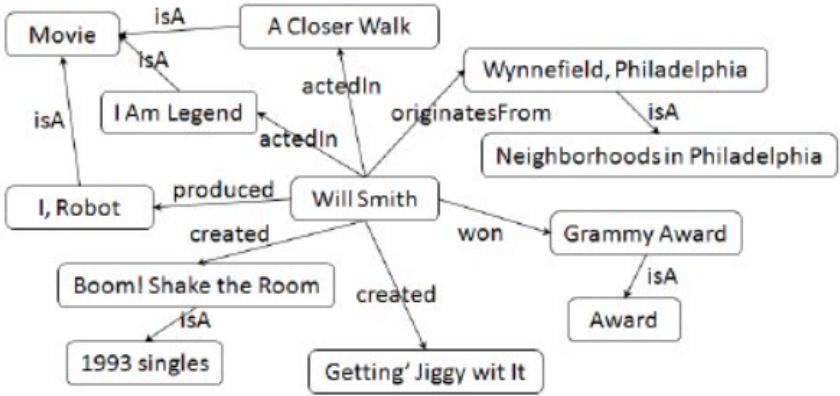
(properSubEvent BurningOfPapalBull-Document
BurningOfPapalBull)

(relationInstanceExists
inputsDestroyed BurningOfPapalBull-Document
(CopyOfConceptualWorkFn
PapalBull-ExcommunicationOfLutherCW)
The thing destroyed is a member of the functionally defined collection "all copies of the conceptual work PapalBull-ExcommunicationOfLuther". The distinction between the conceptual *artifact and the* specific copy *being burned prevents Cyc from concluding that the conceptual work has been utterly destroyed (in the same way that burning a copy of Moby Dick does not destroy the work Moby Dick generally).*

YAGO

- YAGO es una base de alta cobertura que combina Wikipedia y WordNet (actualmente, Schema).
- Similar a Wikipedia en el tamaño pero agrega una taxonomía de conceptos
- Actualmente contiene más de 50 millones de entidades (individuos, organizaciones, lugares, etc.) y más de 2 mil millones de relaciones sobre estas entidades.

Yago - ejemplo



Hay una terna RDF por cada arista del grafo.

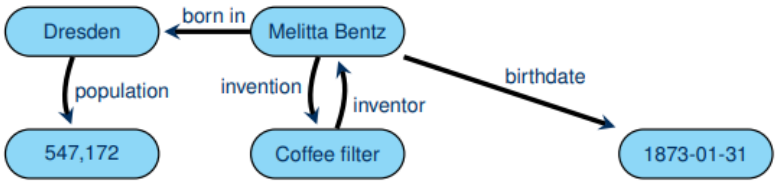
Wikidata

- "Wikidata is a free, collaborative, multilingual, secondary database, collecting structured data to provide support for Wikipedia, Wikimedia Commons, the other wikis of the Wikimedia movement, and to anyone in the world."
- `https://www.wikidata.org/wiki/Wikidata:Main_Page`

Wikidata

- "Wikidata es una Base de Datos secundaria, colaborativa, multilingue y libre ...
- LIBRE Se puede reusar los datos en diferentes escenarios, incluso con fines comerciales.
- COLABORATIVA Los datos se entran y mantienen por editores y bots automáticos.
- MULTILINGUE La edición y el acceso son en una gran variedad de idiomas
- BASE DE DATOS SECUNDARIA. Además de la información, se registran las fuentes y las conexiones con otras bases de datos.

Wikidata - ejemplo



Hay una terna RDF por cada arista del grafo.

KBs, cuadro comparativo

	DBpedia	YAGO	Wikidata	OpenCyc	NELL
Version	2016-04	YAGO3	2016-08-01	2016-09-05	08m.995
# instances	5,109,890	5,130,031	17,581,152	118,125	1,974,297
# axioms	397,831,457	1,435,808,056	1,633,309,138	2,413,894	3,402,971
avg. indegree	13.52	17.44	9.83	10.03	5.33
avg. outdegree	47.55	101.86	41.25	9.23	1.25
# classes	754	576,331	30,765	116,822	290
# relations	3,555	93,659	11,053	165	1,334

https://arxiv.org/pdf/2003.00719.pdf